
Data Science in Action #6

# Proving a reference value that considers all available co-information

Gerhard Svolba
Data Scientist, SAS Austria



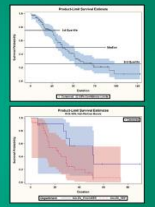
THE POWER TO KNOW.

# Data Science Applications and Case Studies

## Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention

*Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?*

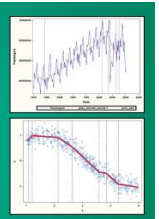


Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

## Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

*Can events and changes in the course over time be automatically detected?*

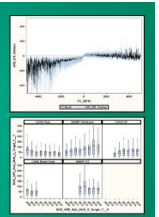


Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

## Data Science in Action: #3

Explaining Forecast Errors and Deviations

*Do the demand planners really improve forecast accuracy with their manual overwrites?*

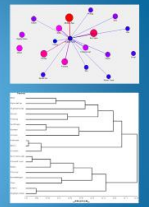


Linear Regression
Quantile Regression
Descriptive Statistics

## Data Science in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

*Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?*

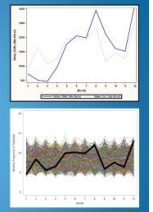


Unsupervised machine learning methods:
association analysis
variable clustering

## Data Science in Action: #5

Checking the Alignment with Predefined Pattern

*Which customers show a behavior that is far from what you expected?*

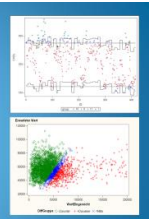


Chi2 independency test
Benford's law
Time Series Similarity

## Data Science in Action: #6

Proving a reference value that considers all available co-information

*Can analytics help me to reduce the "Yes, but … " sentences in my business discussions?*

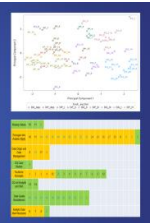


Linear Regression
Decision Trees
Time Series Analysis

## Data Science in Action: #7

Topic Search Documents and Clustering

*Can I automatically find clusters of documents with similar content?*

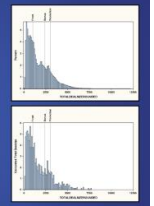


Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

## Data Science in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

*When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?*

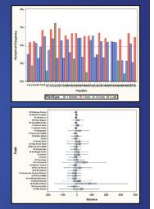


Monte Carlo Simulations
Mathematical Programming

## Data Science in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

*How can you simulate complex environments to get insight in the most frequent processes?*

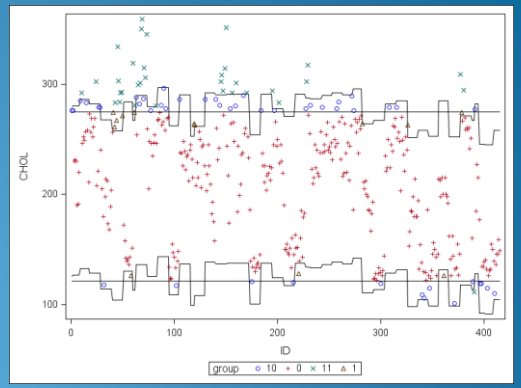


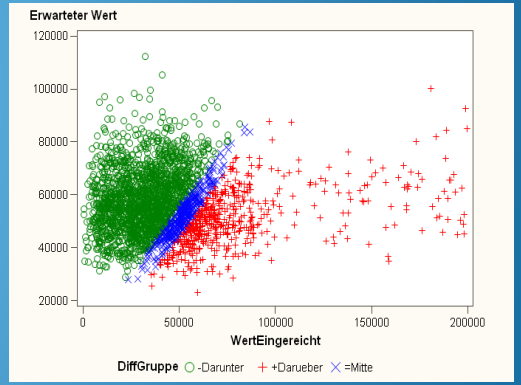
Monte Carlo Simulations

# Data Science in Action: #6

## Proving a reference value that considers all available co-information

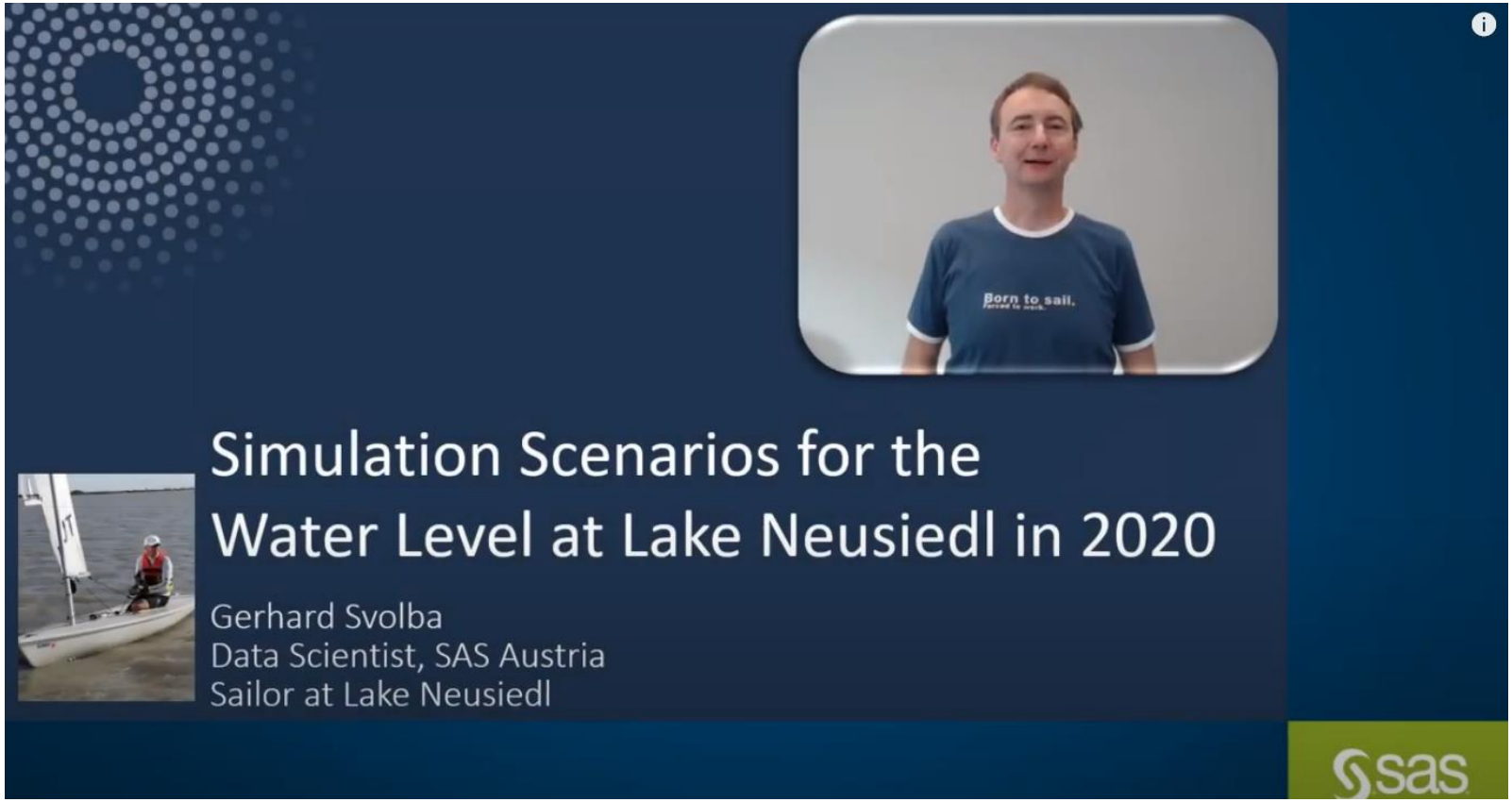
*Can analytics help me to reduce the "Yes, but … " sentences in my business discussions?*



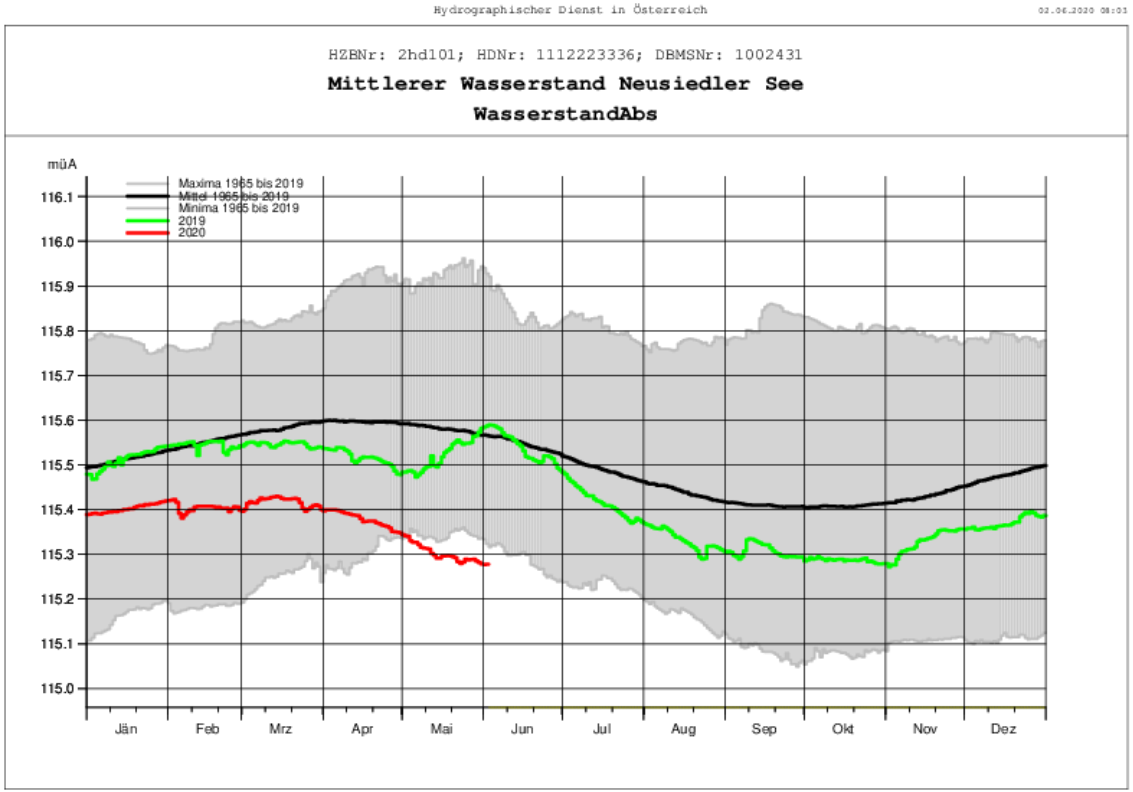



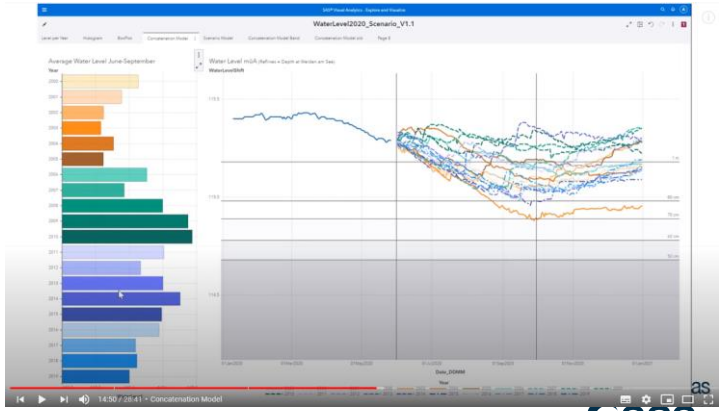

Linear Regression
Decision Trees
Time Series Analysis

# One size does not fit all!
# What is your individual (seasonal) reference value?

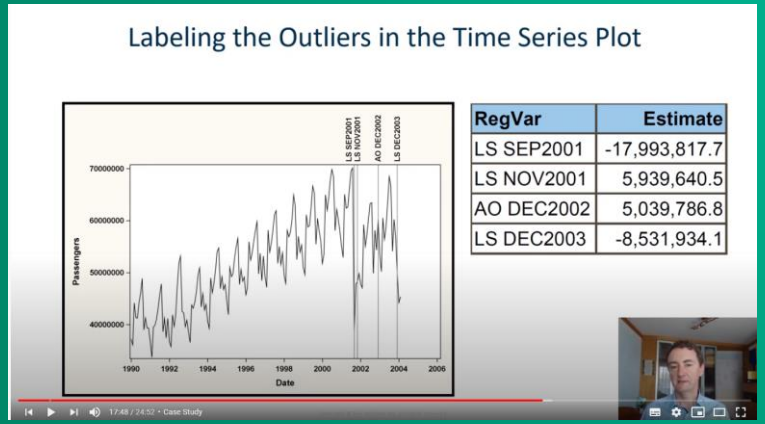


# Considering seasonal variation



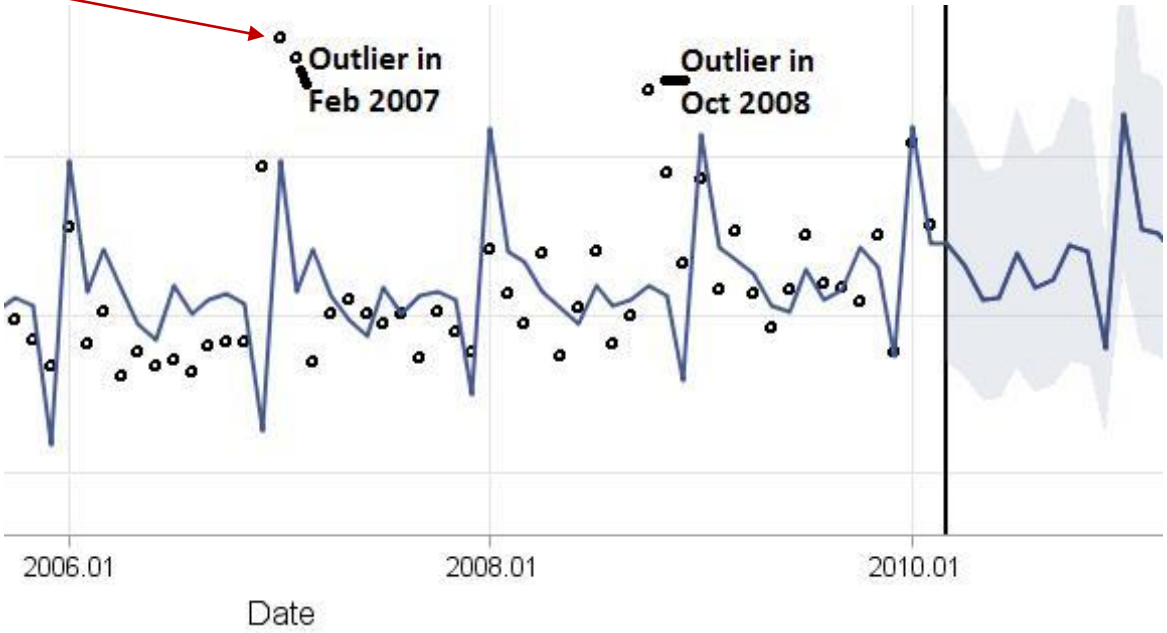
# Does this also work for time series analysis?




| RegVar | Estimate |
|---|---|
| LS SEP2001 | -17,993,817.7 |
| LS NOV2001 | 5,939,640.5 |
| AO DEC2002 | 5,039,786.8 |
| LS DEC2003 | -8,531,934.1 |

# „Yes, but ….
# … in January we always have more events.“

Time Series Model recognizes that this value in January is NOT an outlier

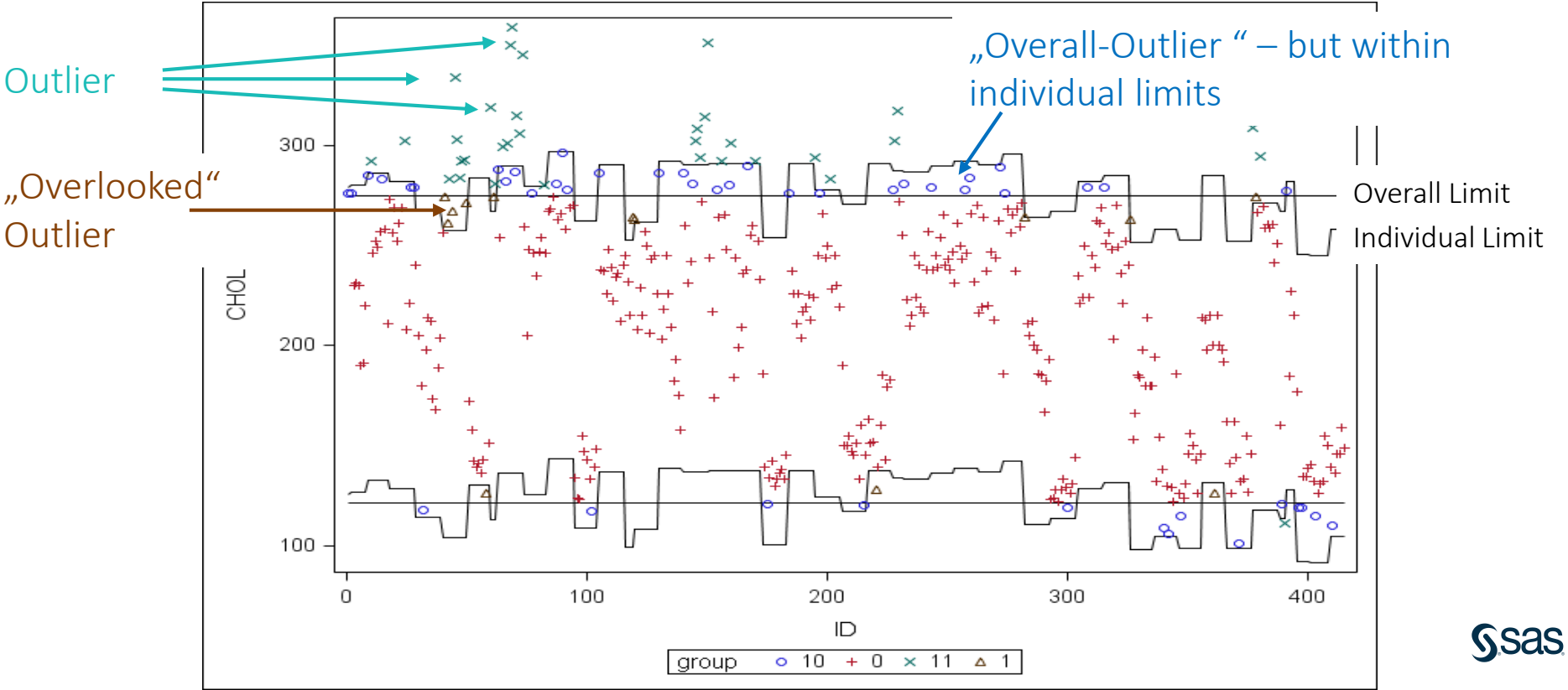


Use a simple regression model to calculate the „expected value“ based on other co-variables

# Calculating the expected cholesterol value based on co-variables

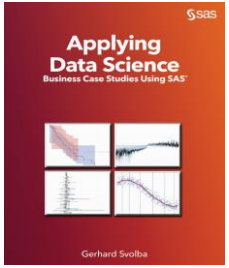
```
proc glmselect data=labor_chol_data;
 class sex centernr stage age_grp weight_grp;
 model chol =  age_grp
               sex
               weight_grp
               centernr
               stage;
 output out=pred_chol p=reference r=residual
            stdi=stdi stdr=stdr stdp=stdp;
run;
```

# *„All values larger than a certain threshold are outliers! - Really?“*



# Analytics and Data Science is there to help you!

- Get a clearer, more objective picture of your data and your analysis subjects
- Get explicit results instead of searching the needle in the haystack
- Make your data talk to you!
- Receive findings automatically instead of manually
- Do it again! – treat models as an asset and repeat your analysis



# Get access to more content:

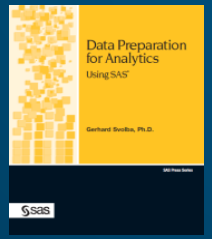
SAS DACH @Youtube: https://www.youtube.com/user/SASsoftwareGermany

Blogs on LinkedIn: https://www.linkedin.com/in/gerhardsvolba/

Twitter: https://twitter.com/gsvolba

Content on Github: https://github.com/gerhard1050

Books @SAS-Press: https://support.sas.com/svolba



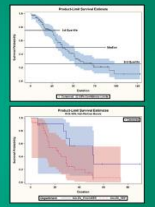


# Data Science Applications and Case Studies

## Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention

*Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?*

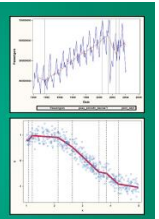


Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

## Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

*Can events and changes in the course over time be automatically detected?*

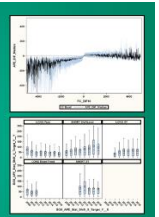


Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

## Data Science in Action: #3

Explaining Forecast Errors and Deviations

*Do the demand planners really improve forecast accuracy with their manual overwrites?*

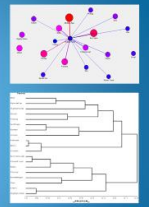


Linear Regression
Quantile Regression
Descriptive Statistics

## Data Science in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

*Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?*

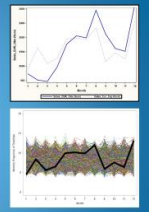


Unsupervised machine learning methods:
association analysis
variable clustering

## Data Science in Action: #5

Checking the Alignment with Predefined Pattern

*Which customers show a behavior that is far from what you expected?*

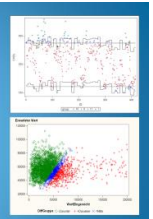


Chi2 independency test
Benford's law
Time Series Similarity

## Data Science in Action: #6

Proving a reference value that considers all available co-information

*Can analytics help me to reduce the "Yes, but … " sentences in my business discussions?*

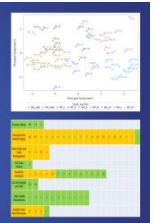


Linear Regression
Decision Trees
Time Series Analysis

## Data Science in Action: #7

Topic Search Documents and Clustering

*Can I automatically find clusters of documents with similar content?*

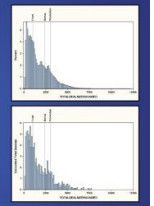


Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

## Data Science in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

*When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?*

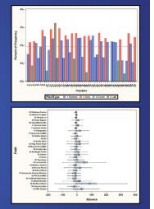


Monte Carlo Simulations
Mathematical Programming

## Data Science in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

*How can you simulate complex environments to get insight in the most frequent processes?*



Monte Carlo Simulations